Dept. of Math. University of Oslo
Statistical Research Report No. 6
ISSN 0806–3842 June 2004

Model-based estimation of transcript concentrations from spotted microarray data.

June 23, 2004

Arnoldo Frigessi[1,2], Mark A. van de Wiel[2,3], Marit Holden[2], Ingrid K. Glad[4] and Heidi Lyng[5]

1. Department of Statistics, Institute of Basic Medical Sciences, University of Oslo, Norway
2. Norwegian Computing Center, Oslo, Norway
3. Present address: Department of Mathematics and Computer Science, Technische Universiteit Eindhoven, The Netherlands
4. Department of Mathematics, University of Oslo, Norway
5. Department of Biophysics, The Norwegian Radium Hospital, Oslo, Norway

Correspondence should be addressed to Heidi Lyng, The Norwegian Radium Hospital, Montebello, 0310 Oslo, Norway. E-mail: helyng@ulrik.uio.no, telephone: +47 22 93 42 58, fax: +47 22 93 42 70).

Key words: gene expression, computational model, Bayesian statistics

Running title: Estimating concentrations in spotted microarrays

Abstract

Much data from spotted microarrays remain unused because obtained with different protocols, platforms or designs, making comparisons across experiments impossible. We have developed a model-based method, which provides absolute transcript levels. Transcript levels are universal, and can be included in further analyses with similar estimates obtained with different techniques in other laboratories. It is a first step both towards genuine meta-analyses, including comparisons across different organisms, and the building of data bases of transcript levels in cells. Our method is based on statistical modelling incorporating all available information about the experiment, from target preparation to image analysis, coherently propagating uncertainties from data to estimates. It requires some genes spotted in replicates, their number being related to the levels of experimental factors included in the model, but not to the number of spotted genes. No uncertainty in the estimates caused by decimated data sets, indirect comparisons, normalisation or imputation of missing values, is introduced, leading to a far more precise analysis of microarray data than provided by conventional methods. Using a flexible Bayesian technique we estimate the highly multivariate joint posterior distribution of all transcripts, which enables extended exploitation of the data. In the present work we apply our method to cervical cancer data. We show that the estimated transcript concentrations are accurate and reproducible, and demonstrate improved statistical tools for selecting genes based on their concentration in highly unbalanced experimental settings.

[Supplemental material is available online at www.genome.org and at http://www.nr.no/pages/samba/area_emr_smbi_transcount].

Efficient production of spotted glass-slide arrays has made the microarray technology to a widespread technique, and improved methods to extract and summarise useful information are needed (Holloway et al., 2002; Butte, 2002; Slonim, 2002). The basic elements are normalised intensity ratios between two biological samples, hybridised together in a single experiment. To allow comparative analysis, the experimental design is transitive, often with a loop or a common reference sample (Churchill, 2002; Yang and Speed, 2003; Townsend, 2003). Such design requirements and the need for stable references impose serious constraints on data use. Methods assessing absolute rather than relative transcript measures would enable integration of data from different sources in global analyses, independent on experimental protocol, design and microarray platform (cDNA and oligonucleotides). Such methods are a principle goal if durable compendia of gene expressions in terms of transcripts per cell, analogous to DNA sequence database information, are to be achieved (Holloway et al., 2002; Moreau et al., 2003).

Extraction of absolute transcript levels from spotted microarray data is complicated due to significant experimental variation and noise originating in the production and hybridisation processes (Butte, 2002; Slonim, 2002; Churchill, 2002). Normalised intensity ratios reduce the influence of systematic effects in the data, though biological information might be lost (Quackenbush, 2002; Kerr et al., 2000). Model-based analysis opens for use of experimental and biological information to increase the accuracy of calculated transcript levels. Linear statistical models have proven successful for identification of differentially expressed genes but absolute transcript levels cannot be obtained (Kerr et al., 2000; Newton et al., 2001; Nygaard et al., 2003).

We have developed a new model based on a radically different principle that enables estimation of absolute transcript levels, thus allowing extended exploitation of microarray data. Experimental information associated with array-, cDNA synthesis-, hybridisation-, and scanning characteristics was incorporated. The model follows the different steps of the microarray experiment. Our method also constitutes an improved analysis tool. We compute the joint posterior distributions of either the absolute or relative transcript levels and reveal dependencies between genes, both within and between individual samples. Uncertainties from sample preparation to imaging have been coherently propagated in a global statistical approach.

Our method was validated on a dataset with known mRNA concentrations. On a second dataset we demonstrate Bayesian analysis. We show that significant results can be obtained from data with limited repetitions. The model can handle experiments based on amplified as well as nonamplified material. The results are based on spotted cDNA microarrays, which

feature particularly large experimental variation, but our model can directly be applied to spotted oligoarrays.

Methods

Principles. The idea is to follow conceptually the mRNA molecules through the microarray experiment, from cDNA synthesis to hybridisation and subsequent washing (Fig. 1). We modelled the process as a stepwise selection, where each molecule had a certain probability of being kept in the experiment. This probability depended on known experimental covariates, like mRNA purity, array, pen, gene, and probe identification, replication, length and quantity. We treated scanning and image analysis as an integral part of the experiment and used associated covariates, such as dye, scanner setting and spot size. Also a scanner and a hybridisation-technique specific characteristic were included: The scanner amplification factor was needed to account for differences in the intensity response among scanner types; the hybridisation factor identified the absolute scale of the estimates. Both were determined in two off-line calibration experiments.

Basic data are the average fluorescence intensities of each spot and their standard deviations, for each experiment. No transformation nor normalisation should be done. Non-transitive data sets are allowed as long as the design includes at least one loop, like a self-self or dye-swap hybridisation. Some genes must be spotted at least in duplicates, their number being independent on the number of genes in the analysis, but related to identifiability of pen effects. In our case, 50 duplicates were enough to identify effects of the six pens used. Experiments with amplified material are handled like those with nonamplified one, but estimates are transformed back to original scale (Nygaard et al., 2003).

We performed Bayesian inference and calculated the posterior joint distribution of all unknown parameters using MCMC (Beaumont and Rannala, 2004). We estimated the number of transcripts for each gene in each sample together with their uncertainty, described by 95% credibility intervals. The posterior joint distribution reflects biological dependency between the number of transcripts, inferred from the data, which cannot be attributed to the experiment. The posterior joint distribution is needed to compute interesting probabilities, such as the probability for a transcript of being among those with highest (or lowest) concentrations.

Covariates. The steps of the microarray experiment were modelled as a binomial selection process, incorporating covariates associated with cDNA synthesis, dye labelling, purification, hybridisation, washing (Fig. 1 and Supplemental Methods 1). The corresponding covariates were array, pen, gene, probe replication (RID), probe identification (PID), probe length, and

probe quantity. Replicated genes had PID and RID effects: PID accounted for different probes, and RID for replications of equal probe. The number of base pairs in the probe sequence was used as probe length. A test slide of each printing series was stained for single stranded DNA by use of SYBR green II (Molecular Probes). The mean spot fluorescence intensities were used as measures of probe quantity. Probe quantity was included since hybridization efficiency of high density probes may be reduced (Peterson et al., 2001).

Covariates associated with scanning were dye, PMT voltage and the scanner amplification factor. The dye covariate represented the dye effect in both labelling and scanning. The amplification factor is a measure of the increase in intensity per unit of increase in PMT voltage. The factor was determined once for each dye and scanner as the slope in log-linear plots of intensity versus PMT voltage (Lyng et al., 2004). A covariates associated with image analysis is the hybridisation factor, used to scale the estimated values to the true number of transcripts. It was determined using two control samples with transcripts at known concentrations, with weighted linear regression of estimates versus true values. Under ordinary stable experimental settings it is sufficient to determine the factor once for each hybridisation method, for example for manual hybridisation and for each type of hybridisation machine.

Statistical Methods. Consider several biological samples. The known quantity of material for sample t on array a is denoted as $q^{t,a}$, for example the weight of mRNA after amplification. For each gene g, let K_g^t denote the unknown number of transcripts per weight unit present in sample t (Fig. 1). Let $L_{j,s}^{t,a}$ be the measured intensity for sample t in pixel j in spot s on array a. The hierarchical, non-linear model that relates these data to the number of transcripts, consists of three layers: (i) a model for the selection process, describing the proportion of target molecules (from the original $q^{t,a} \cdot K_g^t$) that have survived the several steps of the experiment until washing of the hybridised slides; (ii) a model for the scanning process of the hybridised slides; (iii) a model for measurement and residual errors.

In (i), the $q^{t,a} \cdot K_g^t$ molecules undergo a series of processes from cDNA synthesis to hybridisation and washing (Fig. 1). Let n_s^a be the number of pixels in spot s on array a and n_g^a the total number of pixels in all spots related to gene g on array a. After successful cDNA synthesis, labelling and purification, a proportion $c \cdot n_g^a$ of the $q^{t,a} \cdot K_g^t$ molecules candidates to reach the correct spots for hybridisation. Here c is the hybridisation factor per pixel. Each of these $c \cdot n_s^a \cdot q^{t,a} \cdot K_g^t$ molecules has a success probability $p_s^{t,a}$ to hybridise and to remain fixed after subsequent washing, independently of other molecules. This independency corresponds to the usual *probe in excess* assumption. As discussed in Supplemental Methods 1, $p_s^{t,a}$ also accounts for successful cDNA synthesis, dye labelling and purification and it depends on biological and experimental conditions described by covariates. Let $H_s^{t,a}$ be the unknown

number of molecules in sample t that succeeds in hybridising on spot s on array a, resists subsequent washing, thus being available for imaging. Then

$$H_s^{t,a} \sim \text{Binomial}(c \cdot n_s^a \cdot q^{t,a} \cdot K_g^t,\, p_s^{t,a}),$$

where g is the gene spotted in spot s on array a and

$$\begin{aligned} p_s^{t,a} \quad = \quad & \max[1, \exp\{\beta_0 + \beta_e + \beta_a + \beta_p + \beta_g + \beta_{\text{RID}} + \beta_{\text{PID}} \\ & + \beta_l \cdot [\text{probe length}] + \beta_q \cdot [\text{probe quantity}] + \beta_m \cdot [\text{purity}_t]\}]. \end{aligned} \tag{1}$$

The β's represent effects of the various covariates for spot s on array a (β_a array, β_p pen, β_g gene, β_{PID} probe identification, β_{RID} probe replication), [probe length] is the number of base pairs of the probe in spot s, [probe quantity] is the SYBR green intensity, [purity$_t$] is the purity of sample t. $\exp(\beta_0)$ is the global baseline selection probability. When non-transitive data sets are analysed jointly, an effect β_e is required for each transitive subset. Identifiability is assured, see Supplemental Methods 2.

In (ii), the expected scanned intensity on spot s, array a, is modelled as

$$\mu_s^{t,a} = 2^{f_{\text{dye}} \cdot \text{PMT}^{t,a}} H_s^{t,a} \alpha_{\text{dye}}, \tag{2}$$

where PMT^{t,a} is the PMT-voltage used during scanning of sample t on array a, f_{Cy3} or f_{Cy5} are the known scanner amplification factors, while α_{Cy3} and α_{Cy5} are unknown chemical and optical dye effects.

In (iii), we assume for the pixel-wise intensity measurement $L_{j,s}^{t,a}$

$$L_{j,s}^{t,a} = \frac{\mu_s^{t,a}}{n_s^a} + \varepsilon_{j,s}^{t,a}, \tag{3}$$

where $\varepsilon_{j,s}^{t,a}$ is a normally distributed error term with a spot varying variance $(\sigma_s^{t,a})^2$. By conditional independence of the pixel-wise intensities, only the spot-wise mean intensity is required in computations. $\varepsilon_{j,s}^{t,a}$ is estimated directly from the intensities as their sample variance in each spot. Background correction can be included at this level, but was not in our examples.

In the statistical analysis of several arrays and samples, many of the unknown parameters are shared, like array, dye, pen, gene and probe related effects; all data involving sample t contribute information on the unknowns K_g^t. To assure statistical identifiability, some genes must be spotted at least in duplicate. The number of replicated genes is independent on the total number of spotted genes, since replicates are used to estimate the common parameters. The whole data set must include at least one loop, i.e. a self-self array or a dye swap or a longer chain, necessary to identify the relative dye effect $\alpha_{Cy3}/\alpha_{Cy5}$. Beyond this, we do not require a transitive design. To facilitate estimation, the model is reparametrised, so that

the baseline β_0, β_e, β_g, β_m and α_{Cy5} are estimated only on the basis of the variances in the Binomials. Data relative to non-duplicated genes and samples hybridised only once are not used to estimate variances (Supplemental Methods 2). MCMC was implemented to compute the joint and marginal posterior distributions of unknowns of interest (Supplemental Methods 3). The joint distribution describes dependencies between unknowns, for example between K_g^t's for various genes and samples. A priori nothing is assumed on the number of transcripts. The model introduces dependency, through shared experimental factors, so that the quantities $H_s^{t,a}$ are dependent. Dependencies in the data are then attributed backwards in part to this experimental dynamics, and to the posterior joint distribution of the K_g^t's. Estimates of parameters are marginal posterior modes with 95% symmetric credibility intervals.

Materials. Human cDNA microarray slides were printed with 32 pens. Probe length ranged from 525 to more than 2000 base pairs. For validation of our method, 17 DNA control samples were printed in equal amount on six subarrays. We used two control samples, each containing 17 different mRNA sequences, pre-mixed at specific concentrations. 0.5 μl of each sample was used, corresponding to a number of transcripts in the range of $5.8 \times 10^5 - 5.8 \times 10^9$. The concentration ratios achieved when hybridising the two samples together were 1:1, 1:3, 3:1, 1:10, and 10:1 at high and low level concentrations. The labelled samples were hybridized together in a dye-swap design. In a second experiment, two tumour biopsies (A, B) and a reference sample (Ref) of total RNA (Stratagene) were used. The biopsies were from two different locations in a human cervical tumour. Biopsy B was divided into two pieces (B1, B2). Total RNA was isolated (50 to 60 μg) and used to produce labelled cDNA. The samples were hybridized in a loop design (Table 1). RNA purity was optimal and equal for all samples in our experiments and was therefore not used. The slides were imaged at a resolution of 10 μm using an Agilent G2565BA scanner (Agilent Technologies) for slides with control samples and a ScanArray4000 scanner (GSI Lumonics) for slides with biopsies and reference. A laser power of 100% was used. The PMT voltage was adjusted for the red and green channel individually (Lyng et al., 2004). See Supplemental Materials for details.

Results

Validation of the methodology. To validate our method we used dye-swap experiments with control samples at known concentrations. The spot intensities covered the whole detection range, from near background values to saturation. There was a good accordance between the true and estimated number of mRNA molecules (Fig. 2A), but the lowest numbers (below 10^6) were overestimated, consistent with other studies (Held et al., 2003; Dudley et al., 2002;

Hekstra et al., 2003), possibly because low intensity spots had more noise. Background correction improved estimates (data not shown). The uncertainty of our estimates increased for the most abundant molecules with numbers above 10^9. The hybridisation factor was 0.001. Ratios between numbers of molecules per gene in the two samples were also well estimated (Supplemental Figure 7).

We analysed a second, independent experiment, with identical design and protocol (Fig. 2B) using the hybridisation factor 0.001. There was again good accordance between true and estimated values with a systematic underestimation of $\log_{10}$-concentrations by 0.1, small enough not to influence the estimated values significantly, since $\log_{10}$-concentrations were in the range 6 to 10. The hybridisation factor based on the second experiment was 0.0008. The difference between the two hybridisation factors was small.

To illustrate meta-analysis, two non-transitive dye-swap experiments were analysed: samples A and B for the first experiment (Fig. 2A) and samples C and D for the second (Fig. 2B). Here, $A = C, B = D$ and each data set was analysed separately, with hybridisation factor 0.001; the model had no knowledge that samples in the first experiment were repeated in the second one. Since the estimates in Fig. 2A and B were almost equal, we concluded that meta-analysis was successful. The estimated numbers of mRNA molecules for each sample are directly comparable and can be used in further analysis.

Tools for data analysis. In the second experiment, four arrays were hybridised in a loop design with three samples (A, B1, B2) from a human cervical tumour and a reference sample (Ref) (Table 1). It is not clear how to compare optimally the measured intensity ratios with standard methods (Churchill, 2002). We considered 100 genes on 158 spots of each array; 27 genes were duplicated with different probe sequences, 31 genes were duplicated with identical probe sequences, 42 genes were singles. Five different pens were used. This design is unbalanced. Although only a limited number of genes were considered here for illustrative purposes, our method can be equally used for larger data sets of thousands of genes and many samples. To provide concentrations, the estimated numbers of transcripts were related to the known weight of the total RNA.

Estimated concentrations for individual genes were reliable, as pairwise scatterplots (Supplemental Figure 8), and correlations show (Supplemental Table 3). Results were consistent with A, B1 and B2 originating from the same tumour and B1 and B2 originating from the same location within the tumour. We investigated reproducibility of our result by splitting the data into two sets of two arrays each, (Ref-B1, B1-B2) and (B2-A, A-Ref). We analysed these separately, pretending samples were not shared. The estimated numbers of transcripts were very similar for the identical samples, B2 and Ref, showing high reproducibility in our

results (Fig. 3). This similarity supports our claim that estimated numbers of transcripts of different samples can be compared and combined, also when originating from separate experimental schemes, with no transitive design. For example we can compare directly the transcript concentrations in sample A and B1 though the design did not link them transitively.

In experiments based on total RNA it can be investigated if the proportion of mRNA in total RNA is equal for all samples by comparing the sum of all estimated numbers of transcripts in each sample. For the present data with 100 genes, we obtained the sum $5.97 \cdot 10^7$ for sample B1, $5.96 \cdot 10^7$ for B2, $6.17 \cdot 10^7$ for A and $4.49 \cdot 10^7$ for the reference. The similarity of these values for the three tumour samples is consistent with these originating from the same tissue.

We estimated experimental and probe related factors, describing to what extend they influence selection probabilities (Supplemental Table 4). In the second experiment, the four array effects β_a were $-0.54, -0.04, 0.18, 0.40$. This indicated differences in hybridisation efficiency between the four arrays due to non-modelled factors influencing the entire arrays, for example during array production (humidity, temperature). The probe length effect β_l was -0.17, as important as the array effect. The negative sign means that probes with short length have a higher probability to retain molecules for imaging, after hybridisation and washing. Estimated effects can be further used to improve protocols, identifing sources of experimental variation.

Many studies investigate the characteristic gene expressions of a population with a certain trait, using a set of biological samples. We can estimate the characteristic mRNA concentration for each gene of such a population. In the context of the present data set, we first computed the mean of the estimated concentrations of the three tumour samples for each of the 100 genes. The probability densities of mean concentration were often non-normal and skewed (Supplemental Figure 9). Second, we computed for each gene the probability that its mean concentration was among the n highest (Fig. 4). This involves a 100-dimensional integration of the posterior joint distribution performed with MCMC. The steepness of the curve describes the level of concentration of a gene compared to others.

We show two probabilistic gene selection methods: in the first, ranking occurs on the basis of absolute concentrations; the second method requires a threshold on concentrations or on folds of concentration ratios.

First, we evaluated the probability that any single gene in turn had a mean concentration among the highest (or lowest) 10. We then ranked all genes according to this probability (Fig. 5). Low mRNA concentrations are associated with more uncertainty than high ones, resulting in less candidate genes with low concentration (Held et al., 2003). This ranking is independent of any chosen reference sample.

For the second selection method, we considered estimated ratios of mean concentrations in the tumour vs. reference. Suppose we ordered the genes according to a certain criterion, but we only wanted to select those genes that were with high probability at least k-fold expressed. Using the joint distribution, we require that all genes in the selected set are at least k-fold expressed, jointly. Alternatively, we can relax our request and allow m errors (falsely k-fold expressed). Hence, we computed the probability that all but m genes were at least k-fold expressed, and ordered the genes according to the probability that their concentration ratio was among the ten highest or lowest (Fig. 6). Seven genes would be selected, when allowing no errors ($m = 0$) and requiring 2-fold expression ($k = 2$) with 95% joint probability. Alternatively, one may fix the probability (to say 0.95) and the accepted number of errors m and then study the number of selected genes as function of the fold k (Supplemental Figure 10). These plots help choosing candidate genes. The first selection method is useful for concentrations, for which there is no natural cut-off value and we use a probability to rank genes. When natural thresholds are available, like folds of ratios, the second method is useful, since it explicitly controls the joint probability level.

Discussion

We have proposed a new method for estimating precisely the transcript level of individual genes from spotted microarray intensity data and obtained the joint distribution of absolute (or relative) transcript levels, which portraits the dependencies between absolute (or relative) mRNA concentrations. Once the transcript levels have been estimated, radically new analyses are possible, including within sample comparison, merging of data sets with a design lacking transitivity or based on amplified and non-amplified starting materials, cross-platform and cross-species comparisons and more general meta-analysis. This may open for novel approaches in the study of several biological processes, including signal transduction pathways.

Our method is based on four main ideas: we incorporate an extended number of covariates compared to other models (Butte, 2002); we treat unequal number of replicates per gene; we use the binomial process, which better depicts experimental dynamics and allows for estimation of the critical parameters β_0, β_g, and $\alpha_{Cy3/Cy5}$; we avoid normalisation and imputation of missing values and build a bottom-to-top coherent stochastic model, fully propagating uncertainty. The accuracy of our estimates was better than in Dudley et al. (Dudley et al., 2002), especially at medium and low concentrations, and in fact comparable to that achieved from methods based on in situ synthesized arrays (Held et al., 2003; Hekstra et al., 2003), despite this technology uses standardised manufacturing and hybridisation, so that probe

specific biases are highly reproducible and predictable (Li and Wong, 2001).

There are limitations of our methodology. Cross-hybridisation and unspecific binding are not taken into account, and possible splice-variants for some of the genes are not considered. Currently, no analysis tools for microarray data are addressing these aspects. Other covariates could easily be included in our model when available, such as target length and labelling efficiency, probably leading to higher accuracy in the estimates. The MCMC algorithm converges slowly. Results can require up to a few days of computation time.

Few methods estimating absolute transcript levels from spotted microarray data have been developed so far. The method by Dudley et al. (Dudley et al., 2002) requires hybridisation of each sample with a reference of known concentration, imposing serious restrictions. Other methods rely on calibration of each array with additional techniques, such as serial analysis of gene expression (SAGE) (Townsend and Hartl, 2002). The present method is the first quantifying absolute transcript levels from spotted microarray data without the need for calibration of each sample individually. Moreover, it can be used to estimate absolute concentrations from one or multi-color experiments, and it can directly be applied to data from spotted oligoarrays, using base composition of the probes as covariates rather than the probe length. The hierarchical structure of our model enables integration of biological information about the samples, such as patient survival data, and known dependencies between genes, in a coherent Bayesian setting. If the mRNA weight is not available and significant variability in the proportion of mRNA in total RNA is suspected, or if the hybridisation factor is not available, it is possible to scale each sample so that the sum of estimated transcripts are equal. Comparison of such scaled concentrations is still possible between samples, but the interpretation as absolute concentrations is lost.

We estimated concentration ratios more accurately than concentrations themselves, because the uncertainty in the intercept β_0 influences only the absolute numbers and not the ratios. We obtain ratios of concentrations, while usually intensity ratios are compared, whose fold changes can be misleading since they might not correspond to fold changes of actual transcript concentrations (Kerr et al., 2000). In addition our method provides for the first time highly reliable ratios between concentrations of different genes in the same sample.

With our method few constraints are imposed on the experimental design and no normalisation and imputation of missing values is needed. The common reference design requires stable reference samples, uselessly measured many times (Townsend, 2003). A balanced design is required to apply linear mixed effect models in practice (Kerr et al., 2000). Thus our method opens for new possibilities of meta-analyses (Moreau et al., 2003). Such analyses are currently built on top of statistical tests to detect differential expressions (Rhodes et al., 2002;

Choi et al., 2003). Since the result of these tests may depend on experimental protocol and microarray platform, bias may lead to wrong conclusions. With our method, data from different studies can be combined at the basic level of transcript concentrations or concentration ratios, regardless of whether studies use amplified or non-amplified starting material, cDNA or oligonucleotide platforms. Available data can therefore be re-used in new investigations, leading to a better exploitation of the data and more precise results.

In our model, normalisation is performed unsupervised, as in ANOVA based methods (Kerr et al., 2000); we incorporate explicitly more sources of variability, including scanning. Current normalisation methods are often platform dependent and based on hypothesis on the gene expressions difficult to test. Misuse of normalisation is rather common in practice (Yang et al., 2002). The need for balanced designs often leads to discarding genes or requires imputation of missing values. Current methods for imputation fail if the missing mechanism is not at random or if the level of missing exceeds 20% (Troyanskaya et al., 2001).

To identify significantly differentially expressed genes, statistical tests are commonly performed based on normalised intensity ratios (Tusher et al., 2001; Pan, 2003) or on estimated effects (Kerr et al., 2000). Normal distributions cannot be assumed, so that bootstrap and permutation tests are used, requiring a relatively large number of replicates (Yang and Speed, 2003). Our method naturally describes dependencies and does not assume normality. Our ranking schemes and selection criteria are based on the joint distribution of concentrations or concentration ratios. Bayesian assessment of global significance can be easily implemented in our context (Scott and Berger, 2004). Dependencies between genes can be revealed from the joint distribution and graphically represented as a network (Troyanskaya et al., 2003). Since the main experimental factors are corrected for, estimated posterior dependencies can be interpreted as principally of biological origin.

Acknowledgement

We thank L. Holden, E. Hovig, M. Langaas, O. Myklebost, T. Stokke and B. Ylstra for discussions. Financial support was provided by The Norwegian Research Council, The Norwegian Microarray Consortium, The Norwegian Radium Hospital and The Norwegian Cancer Society.

Figure Legends

Figure 1: Illustration of the microarray experiment. The various steps of the experiment and the corresponding covariates used in the model are listed with their symbols. The model consists of three levels: (i) selection, (ii) scanning and (iii) measurement.

In (i), K_g^1 and K_g^2 mRNA molecules for gene g present in sample 1 and 2 undergo a selection process. Each molecule succeeds or fails in each of the experimental steps: cDNA synthesis, dye labelling, purification, hybridisation and washing. Success for each molecule is modelled as a Bernoulli coin toss. The success probability depends on properties of the molecule and of the experiment (covariates). Molecules of the same gene can have different covariates, for example if they hybridise on different spots with different probes. If probe is in excess, molecules can be modelled as independent variables and the number of remaining molecules after each step is Binomially distributed. The probability of successfully passing through the entire experiment is the product of the probabilities of surviving each individual step. Nested Binomial variables are Binomial and the final number of molecules ready for being scanned is Binomial with two parameters: the unknown original number of transcripts per gene in each sample and the selection probability, modelled as in equation (2). Level (ii) describes the translation of the bound molecules remaining after washing ($H_s^{t,a}$, on array a, spot s, for sample $t = 1, 2$) into fluorescence intensities, as in equation (2). Measurement error (iii) of pixelwise intensities $L_{j,s}^{t,a}$ (on array a, pixel j on spot s for sample $t = 1, 2$) is assumed to be normally distributed as in equation (3). This model allows to obtain estimates of absolute concentrations K_g^1 and K_g^2 together with their posterior marginal probability density, as sketched at the bottom.

Figure 2: Validation of the methodology to estimate absolute numbers of transcripts. Control samples with 17 genes of known mRNA concentrations were used, each printed on six spots with six different pens. The inset in panel a shows the posterior probability density of the number of transcripts for a gene with estimated $5.8 \cdot 10^7$ mRNA molecules (mode) and its 95% credibility interval. There was lack of symmetry in the densities. Panel a and b show estimated numbers of mRNA molecules (y-axis) and true ones (x-axis) in $\log_{10}$-scale. Positions on the x-axis are slightly shifted to facilitate visualisation. Diagonal lines are shown; the fit is good when the line passes through the credibility interval. The data in panel A and B are based on two different dye-swap experiments. Analysis of the data in panel B, using the hybridisation factor from the data in panel A (0.001), showed a strong concordance between the two estimates, although the numbers of transcripts were slightly underestimated.

Figure 3: Comparison of the absolute transcript levels in a reference and a human cervical

tumour sample, estimated from two different experiments. The data in Table 1 were split into two sets of two arrays each and analysed separately, pretending no sample was shared. Estimated mRNA concentrations (number of mRNA molecules per μg of total RNA; posterior modes) are plotted for each gene and sample. Diagonal lines are shown. The two independently estimated concentrations Ref_1 and Ref_2 for the reference and B2_1 and B2_2 for sample $B2$ were similar and highly correlated. A small difference was observed for both samples, 0.188 in $\log_{10}$-scale for Ref and 0.231 for $B2$. This difference originated from the uncertainty in the estimation of β_0, a difficult task with just two arrays.

Figure 4: Probability of the four genes (see also Supplemental Figure 9) to be among the n genes with the highest (red curve) or lowest (green curve) mRNA concentration (number of mRNA molecules per μg of total RNA). The plots clearly indicate if a gene is among those with high (gene 46), intermediate (genes 13 and 33) or low (gene 91) concentration.

Figure 5: Probabilities of genes to be among the 10 ones with the highest (red) and lowest (green) mRNA concentrations (number of mRNA molecules per μg of total RNA). See Supplemental Table 2 for gene symbols. There are six genes with probability larger than 0.90 to have mRNA concentration among the 10 highest, and two genes to have concentrations among the 10 lowest ones. The value of the selection probability (here chosen as 0.90) should be as high as possible, but still such that enough genes are selected for the purpose of the study.

Figure 6: For a given group of genes, probability that all but m ratios of mRNA concentrations (number of mRNA molecules per μg of total RNA) in a human cervical tumour vs. reference are at least equal to k. The genes were ordered according to the probability that their ratio was among the ten highest, decreasingly. Gene 90 had highest probability. Up-regulated genes are indicated with "up", down-regulated with "do". We considered then all ordered subsets, following the given ranking: $\{90\}, \{90, 82\}, \{90, 82, 11\}$, and so on. For each such increasing subset of genes we computed the posterior probability that all but m ratios were at least k. Four curves are plotted, for various combinations of m and k. The more genes were included in the selected set, the smaller the probability became. The best set of genes with ratio at least two (k=2) and with at most one error (m=1) with 0.95 joint probability was the set $\{90, 82, 11, 14, 93, 25, 57, 34, 12, 60\}$. The larger the fold k and the smaller the accepted number of errors m, the more rapidly the probabilities decreased.

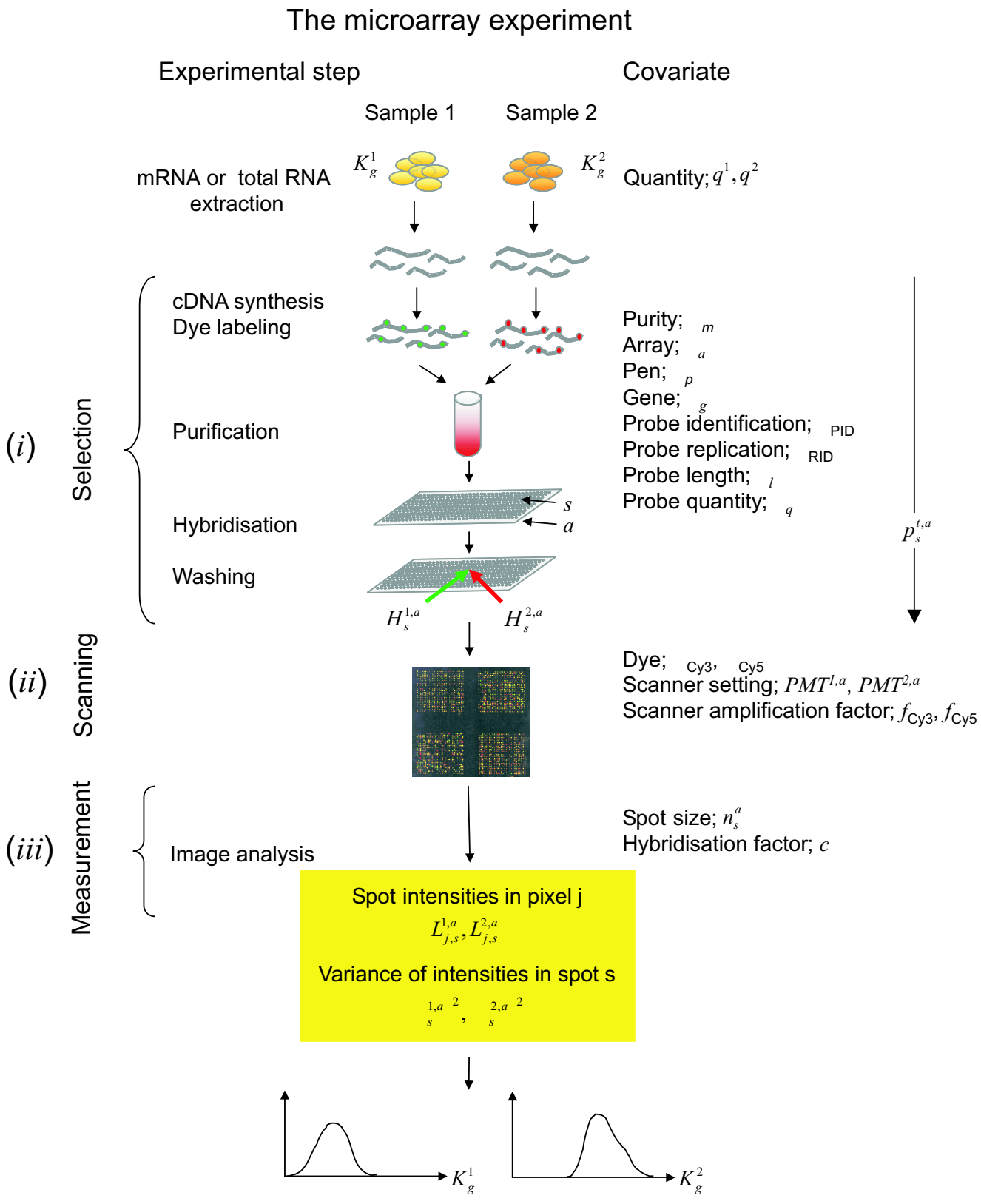
The microarray experiment
Experimental step
Covariate
Sample 1
Sample 2
mRNA or total RNA extraction
K_g^1
K_g^2
Quantity; q^1, q^2
cDNA synthesis
Dye labeling
Purification
Hybridisation
Washing
Purity; $_m$
Array; $_a$
Pen; $_p$
Gene; $_g$
Probe identification; $_{PID}$
Probe replication; $_{RID}$
Probe length; $_l$
Probe quantity; $_q$
$p_s^{t,a}$
s
a
$H_s^{1,a}$
$H_s^{2,a}$
(i) Selection
(ii) Scanning
Dye; $_{Cy3}$, $_{Cy5}$
Scanner setting; $PMT^{1,a}$, $PMT^{2,a}$
Scanner amplification factor; f_{Cy3}, f_{Cy5}
(iii) Measurement
Image analysis
Spot size; n_s^a
Hybridisation factor; c
Spot intensities in pixel j
$L_{j,s}^{1,a}, L_{j,s}^{2,a}$
Variance of intensities in spot s
$_s^{1,a\ 2}$, $_s^{2,a\ 2}$
K_g^1
K_g^2

Figure 1

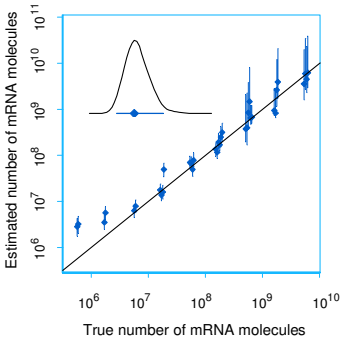

Figure 2a

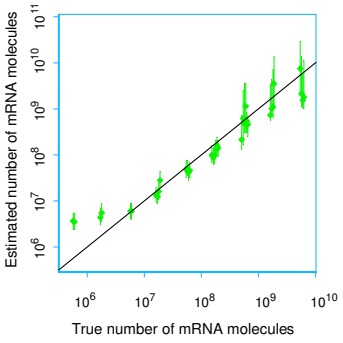

Figure 2b

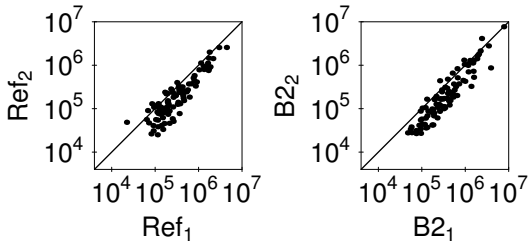

Figure 3

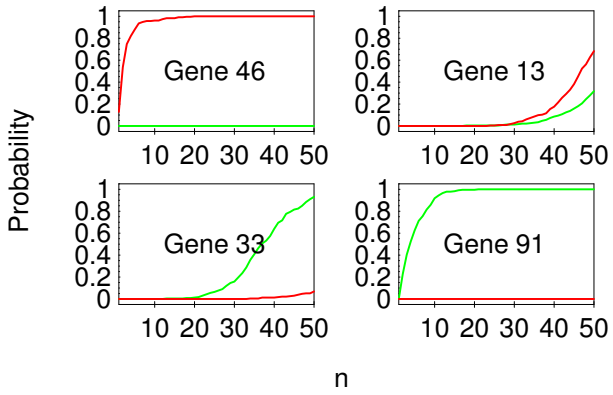
Probability
Gene 46
Gene 13
Gene 33
Gene 91
0
0.2
0.4
0.6
0.8
1
10 20 30 40 50
n

Figure 4

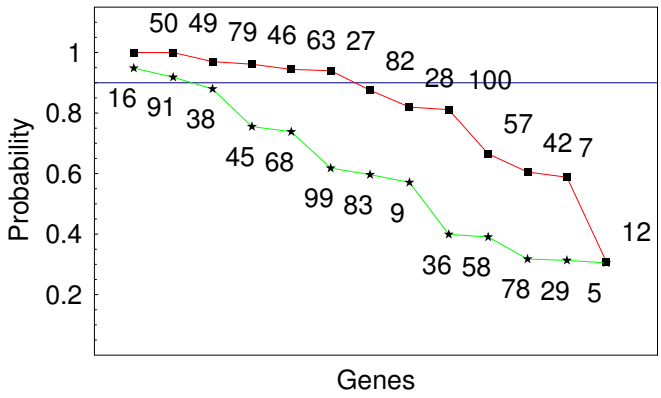
50 49 79 46 63 27 82 28 100 57 42 7 12
16 91 38 45 68 99 83 9 36 58 78 29 5
1
0.8
0.6
0.4
0.2
Probability
Genes

Figure 5

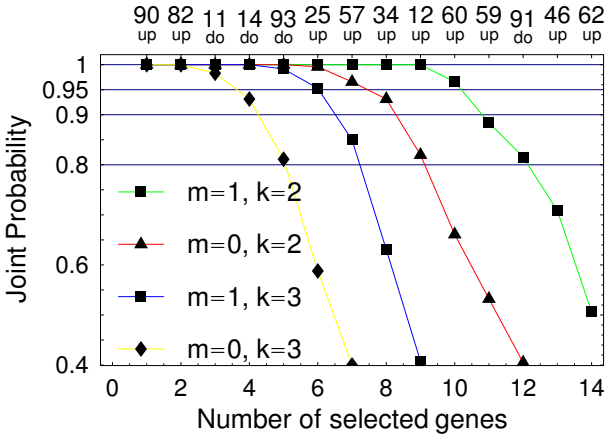
90 up
82 up
11 do
14 do
93 do
25 up
57 up
34 up
12 up
60 up
59 up
91 do
46 up
62 up
1
0.95
0.9
0.8
0.6
0.4
Joint Probability
m=1, k=2
m=0, k=2
m=1, k=3
m=0, k=3
0
2
4
6
8
10
12
14
Number of selected genes

Figure 6

Table 1: Design of human cervical tumour study

Micro-array	Tissue dye Cy5	Tissue dye Cy3
1	Ref	B1
2	B1	B2
3	B2	A
4	A	Ref

Table 1: A, B1 and B2 were samples from the same human cervical tumour. B1 and B2 were derived from the same location within the tumour, A from a different one. Ref was pooled from ten different human cell lines.

References

1. Beaumont M. and Rannala B., 2004. The Bayesian revolution in genetics. *Nat. Rev. Genet.* **5**: 251 – 261.

2. Butte A., 2002. The use and analysis of microarray data. *Nat. Rev. Drug Discov.* **1**: 951–960.

3. Choi J.K., Yu U., Kim S., and Yoo O.J., 2003. Combining multiple microarray studies and modeling interstudy variation. *Bioinformatics* **19 Suppl. 1**: i84–i90.

4. Churchill G., 2002. Fundamentals of experimental design for cDNA microarrays. *Nat. Genet.* **32**: 490–495.

5. Dudley A., Aach J., Steffen M.A., and Church G.M., 2002. Measuring absolute expression with microarrays with calibrated reference sample and an extended signal intensity range. *Proc Natl Acad Sci USA* **99**: 7554–7559.

6. Hekstra D., Taussig A., Magnasco M., and Naef F., 2003. Absolute mRNA concentrations from sequence-specific calibration of oligonucleotide arrays. *Nucleic Acids Res.* **31**: 1962–1968.

7. Held G., Grinstein G., and Tu Y., 2003. Modeling of DNA microarray data by using physical properties of hybridization. *Proc. Natl. Acad. Sci.* **100**: 7575–7580.

8. Holloway A., van Laar R., Tothill R., and Bowtell D., 2002. Options available - from start to finish - for obtaining data from DNA microarrays II. *Nat. Genet. Suppl.* **32**: 481 – 489.

9. Kerr M., Martin M., and Churchill G., 2000. Analysis of variance for gene expression microarray data. *J. Comput. Biol.* **7**: 819–837.

10. Li C. and Wong W., 2001. Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection. *Proc. Natl. Acad. Sci.* **98**: 31–36.

11. Lyng H., Badiee A., Svendsrud D.H., Hovig E., Myklebost O., and Stokke T., 2004. Profound influence of non-linearity in microarray scanners on gene expression ratios: Analysis and procedure for correction. *BMC Genomics* **5:10**.

12. Moreau Y., Aerts S., De Moor B., De Stooper B., and Dabrowski M., 2003. Comparison and meta-analysis of microarray data: from the bench to the computer desk. *Trends Genet.* **19**: 570–577.

13. Newton M., Kendziorsky C., and Richmond C.e., 2001. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data. *J. Comp. Biol.* **8**: 37–52.

14. Nygaard V., Løland A., Holden M., Langaas M., Rue H., Liu F., Myklebost O., Fodstad Ø., Hovig E., and Smith-Sørensen B., 2003. Effects of mRNA amplification on gene expression ratios in cDNA experiments estimated by analysis of variance. *BMC Genomics* **4:11**.

15. Pan W., 2003. On the use of permutation in and the performance of a class of nonparametric methods to detect differential gene expression. *Bioinformatics* **19**: 1333–1340.

16. Peterson A., Heaton R., and Georgiadis R., 2001. The effect of surface probe density on DNA hybridization. *Nucleic Acids Res.* **29**: 5163 – 5168.

17. Quackenbush J., 2002. Microarray data normalisation and transformation. *Nat. Genet.* **32**: 496–501.

18. Rhodes D.R., Barrette T.R., Rubin M.A., Ghosh D., and Chinnaiyan A.M., 2002. Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway disregulation in prostate cancer. *Cancer Res* **62**: 4427–4433.

19. Scott J. and Berger J.O., 2004. An exploration of aspects of Bayesian multiple testing. Technical Report 2003, Duke University, www.isds.duke.edu/~berger/papers/multcomp.pdf.

20. Slonim D., 2002. From patterns to pathways: gene expression data analysis comes to age. *Nat. Genet.* **32**: 502–508.

21. Townsend J., 2003. Multifactorial experimental design and the transitivity of ratios with spotted DNA microarrays. *BMC Genomics* **4:41**.

22. Townsend J. and Hartl D., 2002. Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biol.* **3**: research0071.1–0071.16.

23. Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D., and Altman R., 2001. Missing value estimation methods for dna microarrays. *Bioinformatics* **17**: 520–525.

24. Troyanskaya O., Dolinski K., Owen A.B., Altman R.B., and Botstein D.A., 2003. Bayesian framework for combining heterogeneous data sources for gene function prediction (in Saccharomyces cerevisiae). *Proc. Natl. Acad. Sci.* **100**: 8348 – 8353.

25. Tusher V., Tibshirani R., and Chu G., 2001. Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci.* **98**: 5116–5121.

26. Yang Y., Dudoit S., Luu P., Lin D.M., Peng V., Ngai J., and Speed T., 2002. Normalisation for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res* **30**.

27. Yang Y. and Speed T., 2003. *Design and analysis of comparative microarray experiments*, 35–92. Chapman and Hall.

Web Site References

http://www.nr.no/pages/samba/area_emr_smbi_transcount, contains data and software used in this paper, available for download.